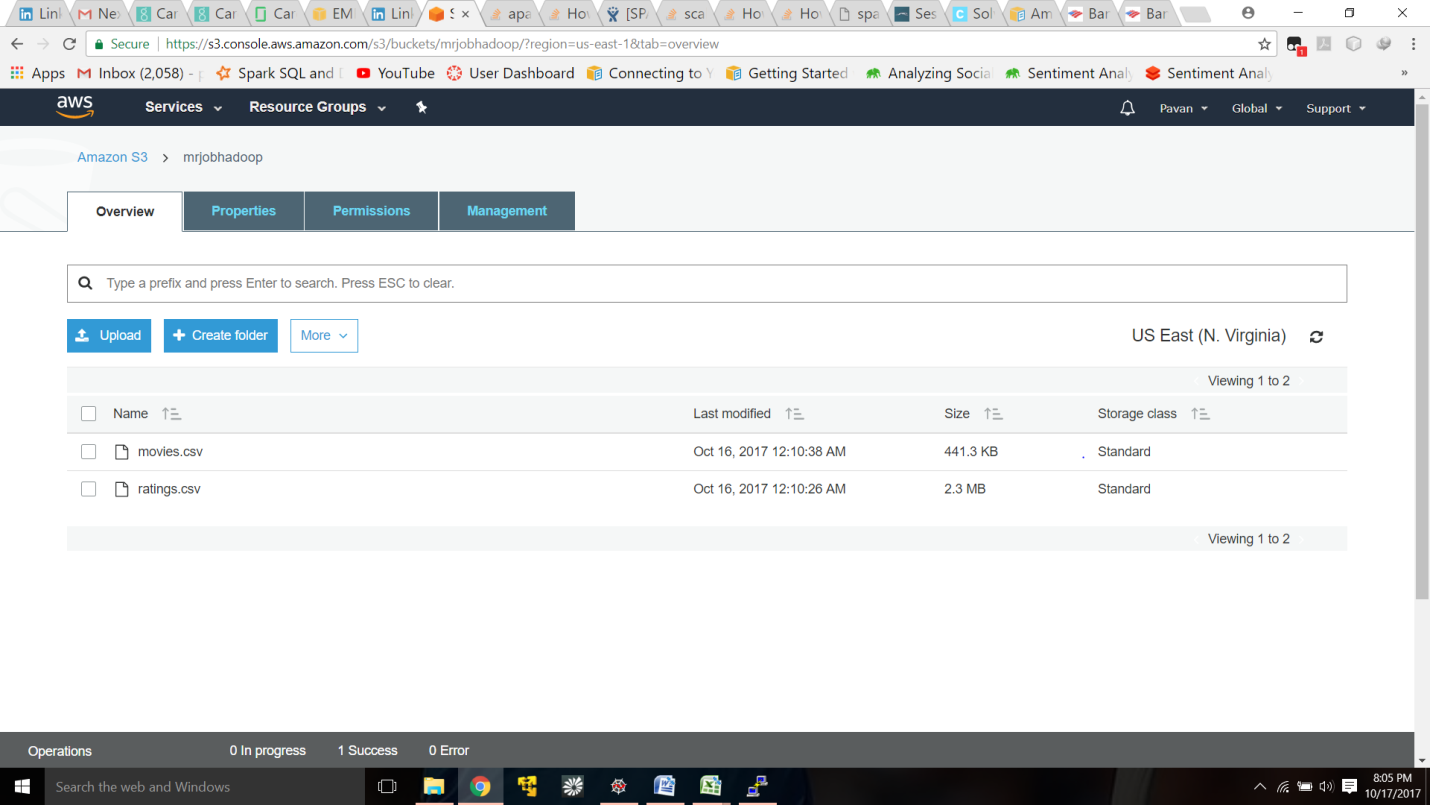
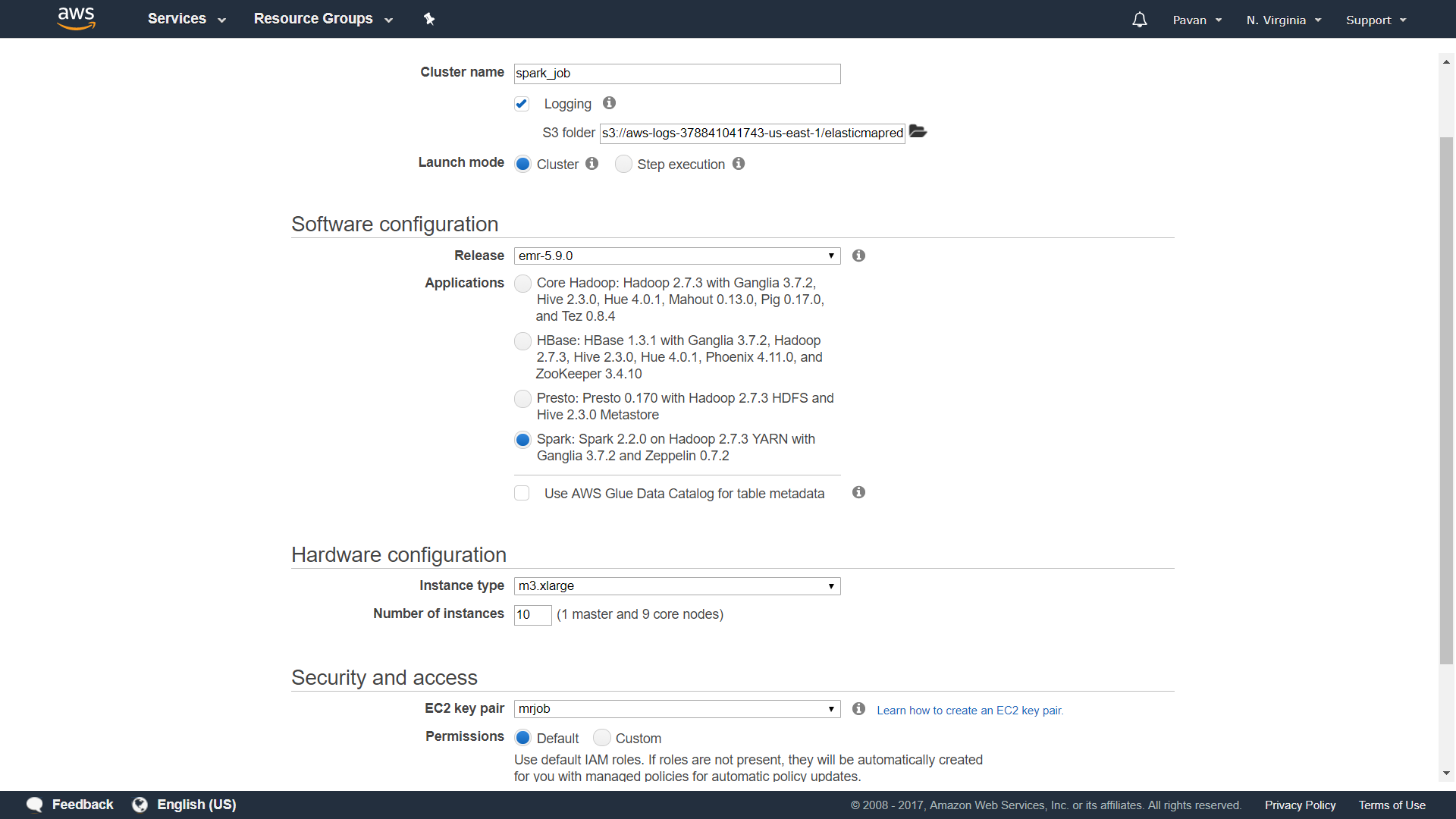
**Project Documentation**

**Step 1:**

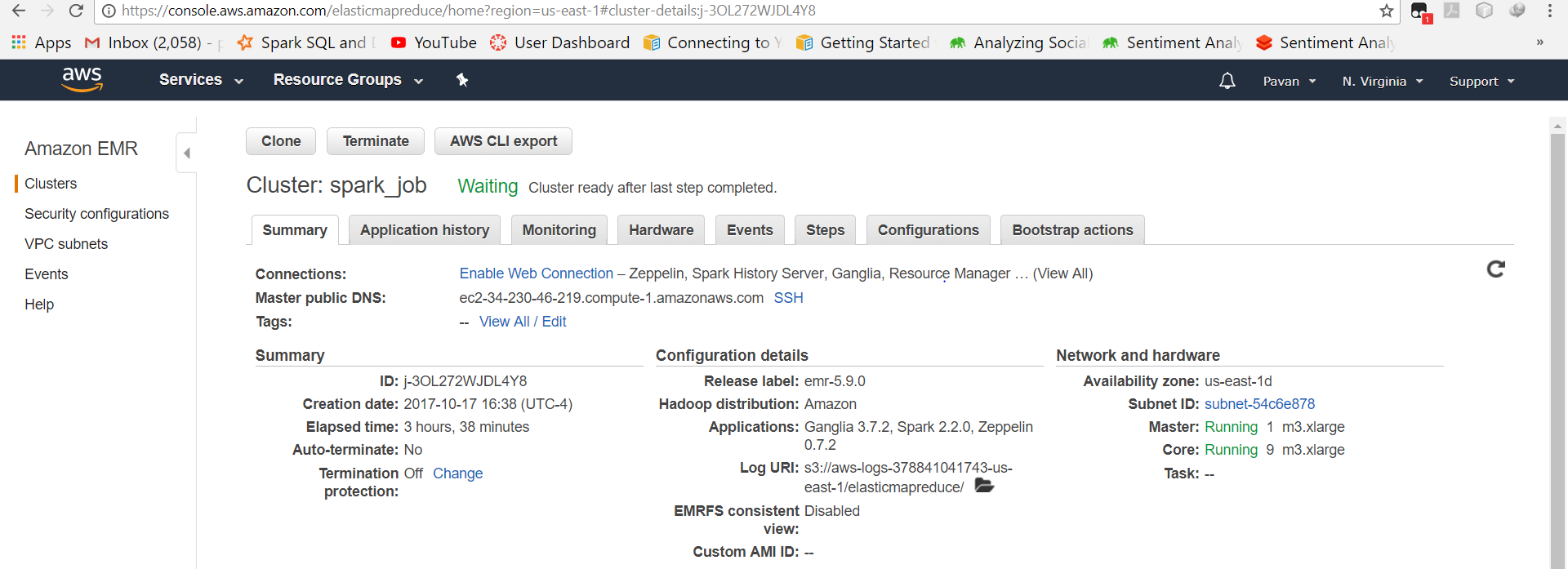
Now I Created a bucket in s3. Copied the required input files into the bucket. As shown below.



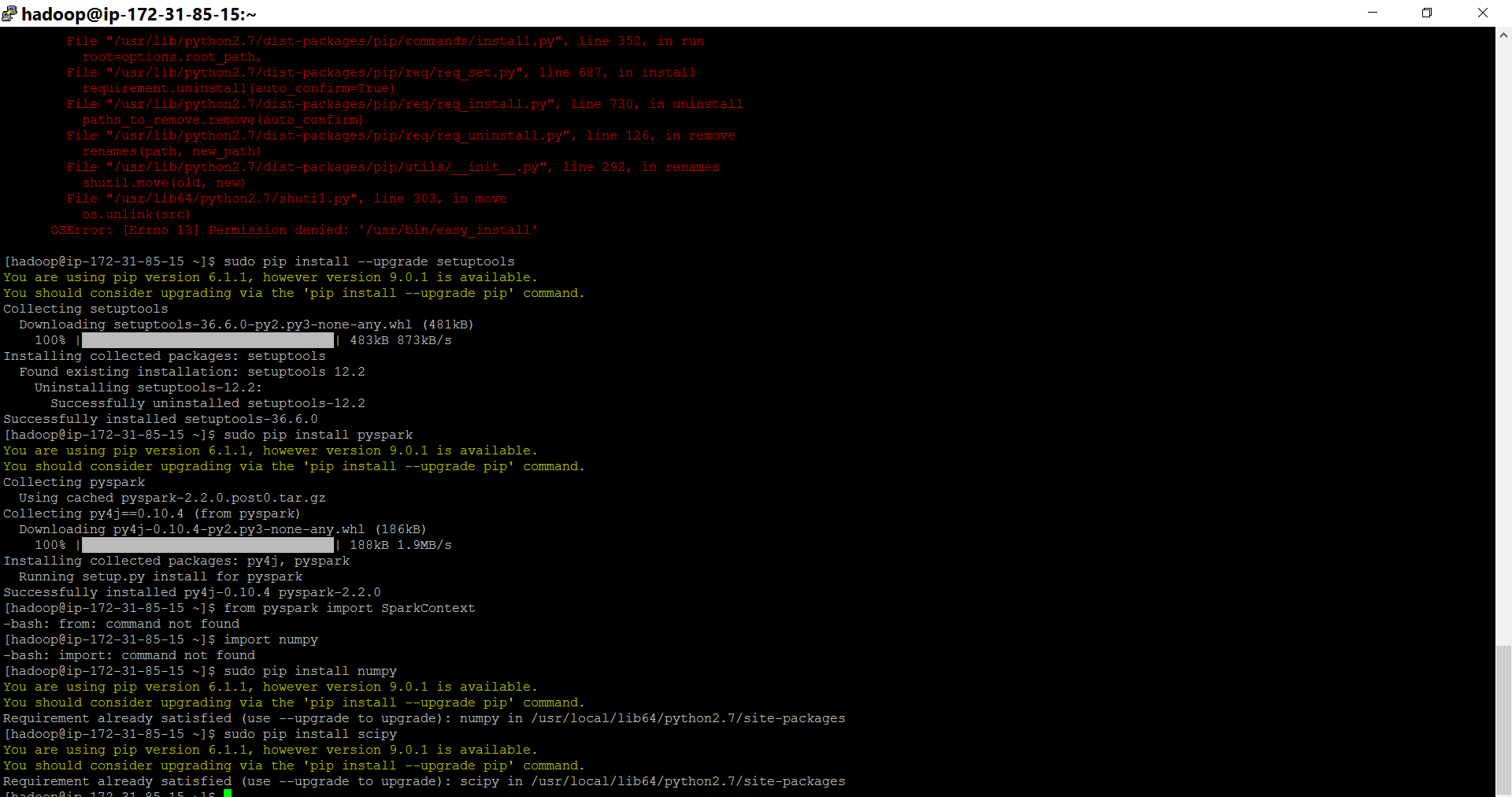
**Step 2:** Created the cluster using AWS EMR module. while creating I selected spark 2.2.0 application and m3x.large as instance type and I selected 10 instances to form a cluster as shown below.

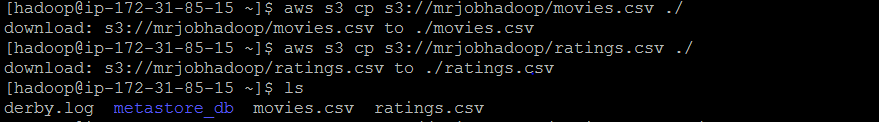


Now the cluster is in running state.

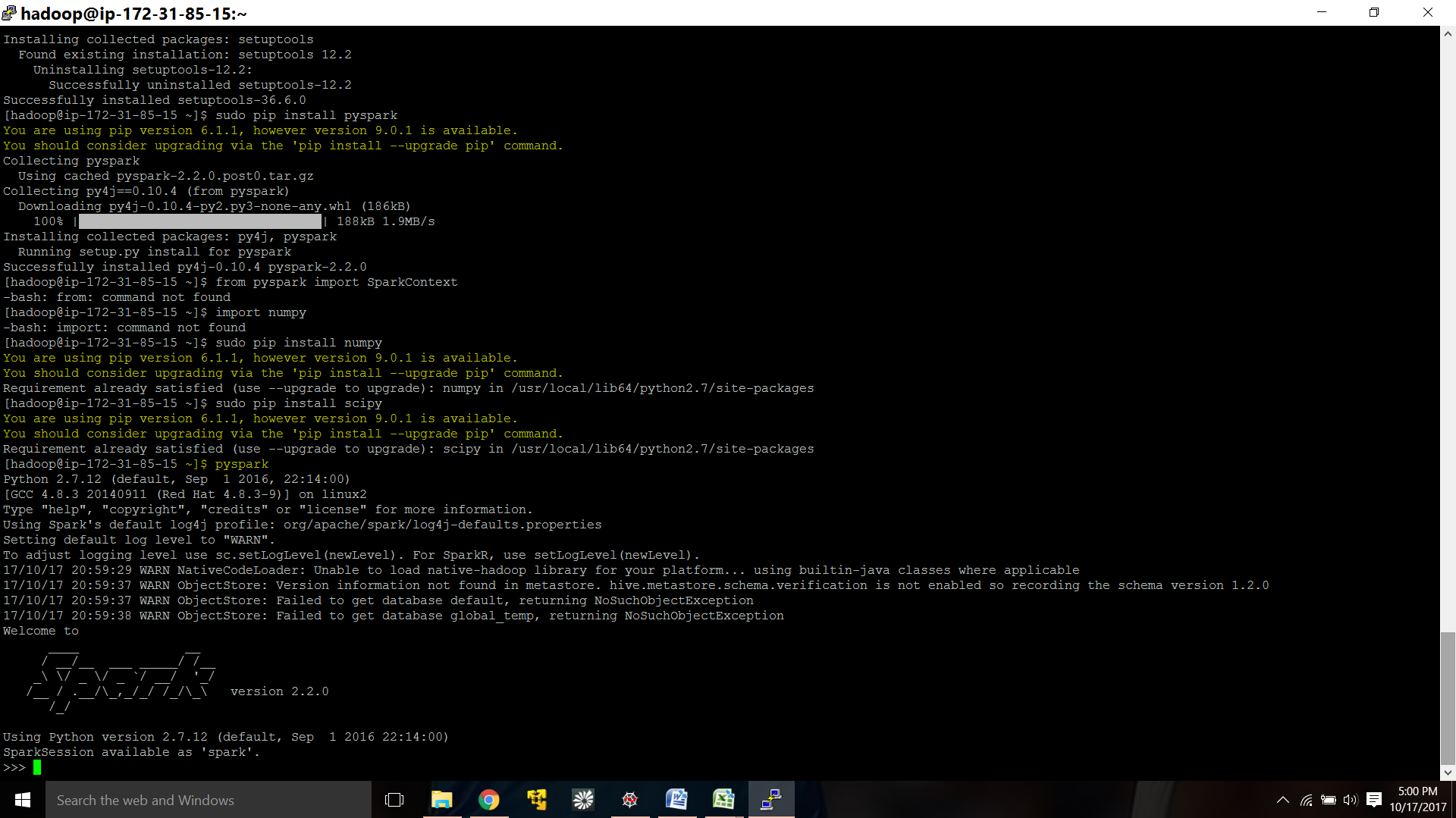


I connect to the master node of cluster using putty and I Installed the required packages in ec2 instance. Please find the below screenshot.

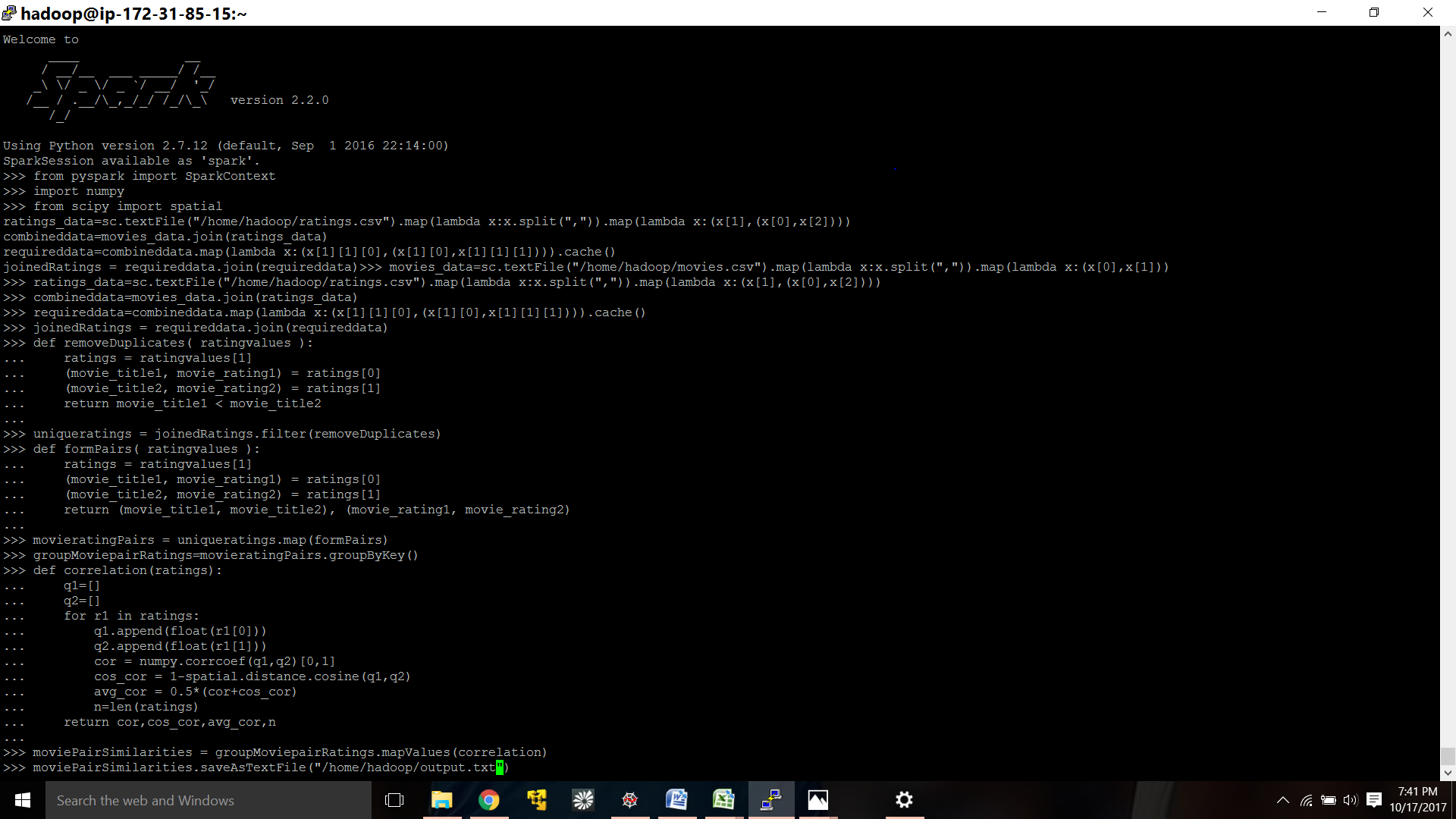


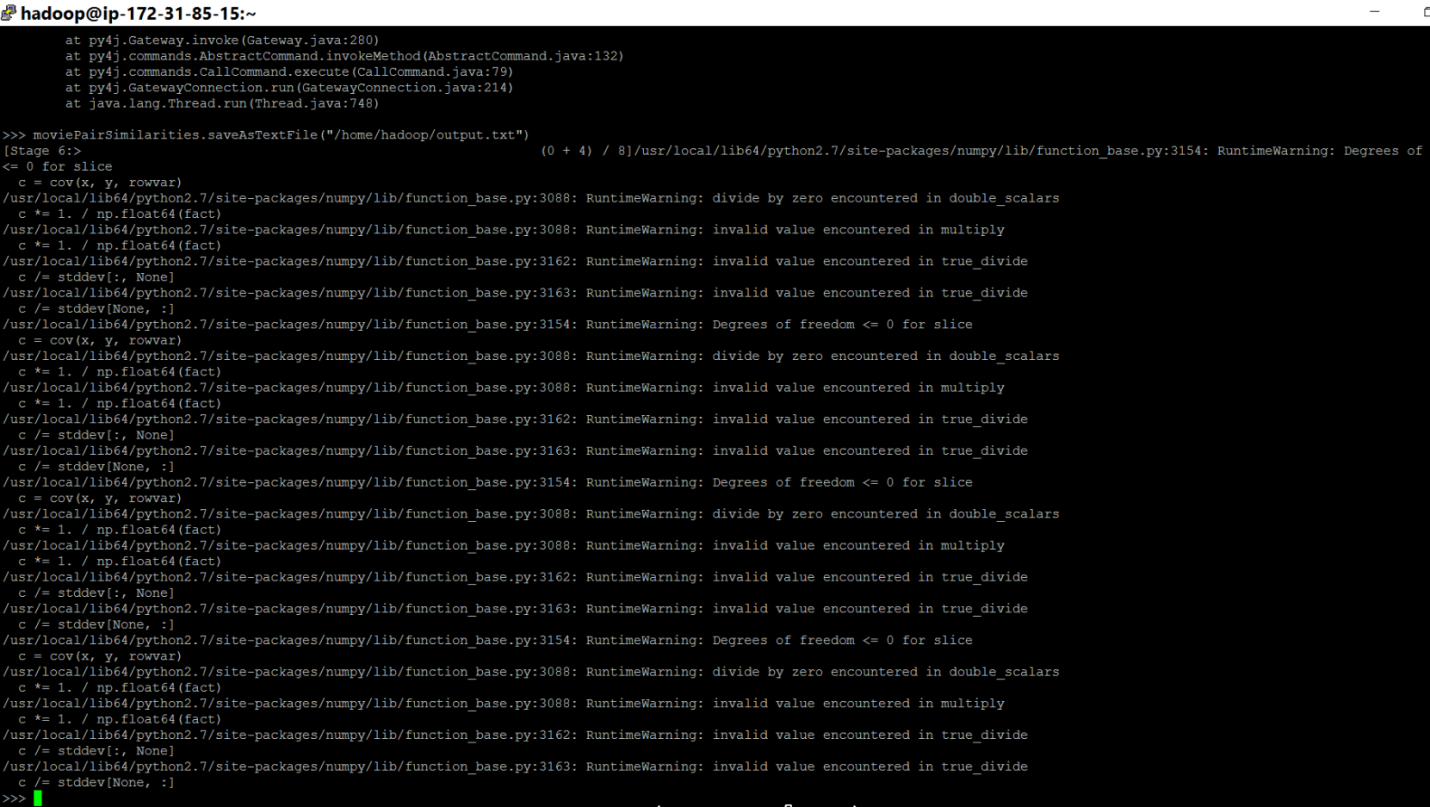
Copied the required documents in to instance as shown below . 

Started the spark shell in the instance using **pyspark** command as shown below.



Now I executed the code line by line in the shell. Please find the below screenshot.





Please find the output.txt file in the instance as shown below.

